研究指導　中澤 真 准教授

感情辞書の効率的な作成手法の提案 —食べログの口コミを題材にして—

北原　広也

1. はじめに

近年ソーシャルメディアの普及によって，ソーシャルネットワークサービスには毎日大量の文章が書きこまれている．その中には，マーケティングにつながるユーザの感情・評判などを知ることができる重要な手掛かりとなる情報が多くある．そのため，これらの情報を収集し，分析をすることで企業側はマーケティングへとつなげることができる．しかし，ユーザの感情を分析する際，感情を表現している単語を抽出しなければならない[1][2]．また，正確に感情を取り出すためには正しい答え，すなわち辞書が必要である．

三和ら[3]は，ランダムに集めたツイートから文章を評価する際の決め手となった単語をポジティブ・ネガティブに分け，「極性値のある単語の辞書」と「強弱表現などの辞書」との2つに分けた自作辞書に手動で単語を追加して，極性値の高い単語や強弱表現のある単語を用いて感情分析する手法について述べている．しかし，感情分析において人間があらかじめ辞書を作成することは非効率であり，その上適切な言葉を選ぶことができていない可能性もある．

そこで本研究ではキーワードの出現頻度に注目して，ポジティブ・ネガティブ，それぞれの感情を判別するのに適したキーワードを抽出し，感情辞書を自動作成する手法を提案する．

2. 感情分析の現状

ソーシャルメディアの普及で，それらに投稿されたユーザの書き込みや，その中に含まれる本音などが重要視されるようになった．そのことから，マーケティングにおいて感情分析・評判分析をすることは，消費者のニーズに柔軟に応えるために重要である[2][3]．しかし，正確に感情分析・評判分析ができたという研究は，感情語の定義が明確にないことから存在しない[4][5]．

感情分析ではポジティブ・ネガティブな感情を判別するキーワードの辞書をあらかじめ登録する方法が一般的である．しかし，これは非常に手間のかかる作業であり，効率的ではない．その上適切な言葉を選ぶことができていない可能性も考えられる．そこで，感情辞書を自動で作成することで，手動で感情辞書を作成するよりもより適切な言葉を選ぶことができると考えた．

3. 感情辞書作成のための口コミ分析

本研究では，ポジティブ・ネガティブそれぞれの文章の評価における特徴語を見出し，そこから辞書を自動作成する手法を提案する．提案手法では，五段階評価で4以上の口コミの集合において出現率が高く，かつ2以下の評価の口コミ集合において出現率の低かった単語をポジティブな特徴語とする．逆に，2以下の評価の口コミ集合において出現率が高く，かつ4以上の評価の口コミ集合において出現率が低かった単語をネガティブな特徴語とする．

利用するデータとして，「食べログ」に投稿されている口コミをランダムサンプリングで80件選び出す．評価4以上の17件の口コミをポジティブな文章集合とし，評価2以下の27件の口コミをネガティブな文章集合とした．そして，それぞれの文章集合に出現した単語を細かく調査するために，RMeCabを用いて形態素解析[1]を行い，出現頻度を算出する[2][6]．

一般的に形態素解析した単語の分析をする際は名詞・動詞・形容詞を主に扱う．そこで本研究では，ポジティブ・ネガティブそれぞれの評価の文章を形態素解析した後，出現頻度を算出した．そしてその単語を形容詞・動詞・名詞に分類し，それぞれの出現頻度上位の単語を特徴語の候補とし，感情辞書作成のために用いる．

4. 分析結果と検証

4.1 形容詞における特徴語の抽出

ポジティブな文章集合において出現率の高かった形容詞の上位10位を表1に示した．また，ネガティブな

1 文章を意味のある単語に区切り，辞書を利用して品詞や内容を判別すること．コンピュータによる自然言語処理技術の一つ．

文章集合において出現率の高かった形容詞の上位10位を表2に示した．ここで表中の出現率は，単語の出現頻度をそれぞれの文章集合に含まれる口コミ数で除したものである．表を見るとポジティブな文章集合において「美味しい」，「良い」などの形容詞の出現頻度が高いことがわかるが，これらの単語はネガティブな文章集合でも出現率が高い．このような両方の集合に高い出現率で現れる単語は，それぞれの文章集合固有の単語とはいえず特徴語としては適切ではない．

そこで，2つの口コミの集合に現れている形容詞で，出現率の順位差が20位以上あったものを形容詞の特徴語とし表中で網掛け表示にした．なお，「嬉しい」，「しょっぱい」といったポジティブ・ネガティブいずれかの文章にしか現れなかった単語も，形容詞における特徴語としている．

この結果から形容詞における二つの文章集合の特徴語を選んだ場合，ネガティブな文章における特徴語はわずか2語となり非常に少ない．そこで，対象となる形容詞を出現率の上位20位まで拡げることで，ポジティブ・ネガティブを表す形容詞の特徴語を追加することにした．このようにして抽出された形容詞の特徴語を表3に示す．

表1:ポジティブな文章集合において
出現率の高い形容詞の上位10位の一覧

星4以上順位	星2以下順位	形容詞	4以上での出現	星2以下での出現率
1	2	美味しい	124%	104%
2	1	ない	106%	115%
3	4	良い	100%	59%
4	8	多い	59%	33%
5	42	旨い	47%	7%
6	該当なし	まぶしい	35%	0%
7	85	濃い	35%	4%
8	3	無い	35%	100%
9	該当なし	嬉しい	29%	0%
10	5	高い	29%	37%

表2:ネガティブな文章集合において
出現率の高い形容詞の上位10位の一覧

星2以下順位	星4以上順位	形容詞	2以下での出現	星4以上での出現率
1	2	ない	115%	106%
2	1	美味しい	104%	124%
3	8	無い	100%	35%
4	3	良い	59%	100%
5	10	高い	37%	29%
6	11	いい	33%	24%
7	14	安い	33%	18%
8	4	多い	33%	59%
9	該当なし	しょっぱい	26%	0%
10	30	悪い	26%	6%

表3:形容詞の特徴語の一覧

ネガティブ		ポジティブ	
星2以下順位	形容詞	星4以上順位	形容詞
9	しょっぱい	5	旨い
10	悪い	6	まぶしい
11	酷い	7	濃い
12	イイ	9	嬉しい
13	大きい	12	すごい

4.2　動詞における特徴語の抽出

形容詞と同様にポジティブな口コミの文章集合において出現率の高かった動詞の上位10位を表4に示した．また，ネガティブな文章集合において出現率の高かった動詞の上位10位を表5に示す．どちらの文章集合にも高い出現率で現れる動詞は，「する」「ある」「食べる」などグルメサイトの口コミとしてはポジティブ・ネガティブに関係なく使用されるものであった．

次に形容詞と同様に2つの口コミ集合の出現率順位差が大きいものを特徴語とし，表中に網掛けで表示した．なお，動詞では順位差が50位以上のものを特徴語としている．

この結果から動詞における二つの文章集合の特徴語を選んだ場合，ネガティブな文章集合では「合う」「見える」の2語，ポジティブな文章では「言う」の1語であり非常に少ない．そこで，対象となる動詞を出現率の上位40位まで拡げることで，ポジティブ・ネガティブを表す動詞の特徴語を追加することにした．このようにして抽出された動詞の特徴語を表6に示す．

表4:ポジティブな文章集合において
出現率の高い動詞の上位10位の一覧

星4以上順位	星2以下順位	動詞	4以上での出現	星2以下での出現率
1	1	する	529%	937%
2	2	ある	176%	315%
3	3	食べる	141%	233%
4	5	思う	106%	159%
5	4	なる	94%	167%
6	8	入る	71%	81%
7	65	合う	53%	11%
8	7	行く	41%	85%
9	101	見える	35%	7%
10	16	いる	29%	44%

表5:ネガティブな文章集合において
出現率の高い動詞の上位10位の一覧

星2以下順位	星4以上順位	動詞	2以下での出現	星4以上での出現率
1	1	する	937%	529%
2	2	ある	315%	176%
3	3	食べる	233%	141%
4	5	なる	167%	94%
5	4	思う	159%	106%
6	118	言う	130%	6%
7	8	行く	85%	41%
8	6	入る	81%	71%
9	42	感じる	63%	12%
10	21	できる	59%	18%

表6:動詞の特徴語の一覧

ネガティブ		ポジティブ	
星2以下順位	動詞	星4以上順位	動詞
6	言う	7	合う
13	待つ	9	見える
14	来る	11	盛る
18	空く	18	並ぶ
26	出す	28	買う
		32	そそる

4.3　名詞における特徴語の抽出

名詞についてもポジティブ・ネガティブそれぞれの文章集合において出現率の高い単語を抽出した．その結果を表7，表8に示す．どちらの文章集合にも高い出現率で現れる名詞は，「店」「味」などのそれ自身には感情の情報が含まれないものであった．

この結果から2つの口コミ集合の出現率順位差が，大きい単語を抽出して表中で網掛け表示した．この際に，食材や料理名などの一般名詞を除外し，「満足」「感動」などのサ変接続の名詞を中心に選んだ．なお，名詞では順位差が60位以上のものを特徴語としている．また，出現率の下限を拡げて対象となる特徴語を増やした．この結果を表9に示す．

表7:ポジティブな文章集合において
出現率の高い名詞の上位10位の一覧

星4以上順位	星2以下順位	名詞	4以上での出現	星2以下での出現率
1	1	店	135%	348%
2	2	味	124%	204%
3	15	スープ	82%	59%
4	29	メニュー	65%	41%
5	134	満足	59%	15%
6	該当なし	キクラゲ	53%	0%
7	20	麺	53%	52%
8	387	ネギ	47%	7%
9	17	注文	41%	56%
10	104	ご飯	41%	19%

表8:ネガティブな文章集合において
出現率の高い名詞の上位10位の一覧

星2以下順位	星4以上順位	名詞	2以下での出現	星4以上での出現率
1	1	店	348%	135%
2	2	味	204%	124%
3	71	評価	115%	18%
4	55	客	100%	24%
5	72	訪問	93%	18%
6	17	料理	89%	35%
7	35	そば	78%	29%
8	11	ラーメン	78%	41%
9	20	人	78%	35%
10	45	雰囲気	74%	29%

表9:名詞の特徴語の一覧

ネガティブ		ポジティブ	
星2以下順位	名詞	星4以上順位	名詞
5	評価	5	満足
7	訪問	32	感動
19	値段	57	最高
21	接客	90	常連
27	サービス	98	満席
48	残念	100	素直

表3，表6，表9から最終的にポジティブな特徴語として選び出した単語を表10に示し，同様に最終的にネガティブな特徴語として選び出した単語を表11に示す．

表10:ポジティブな特徴語

旨い	まぶしい	濃い	嬉しい	すごい
合う	見える	盛る	並ぶ	買う
そそる	満足	感動	最高	常連
満席	素直			

表11:ネガティブな特徴語

しょっぱい	悪い	酷い	大きい	薄い
イイ	言う	待つ	来る	空く
出す	評価	訪問	値段	接客
サービス	残念			

5.　感情辞書への登録語としての適正検証

先の分析結果から抽出したポジティブ・ネガティブの特徴語が，感情辞書に登録する単語として適正であるかどうかを確認するための評価検証をする．形態素解析に使用した評価4以上の口コミ17件と，評価2以下の口コミ27件に対し，抽出した特徴語によってこれらの文章集合を分類できるかを確かめる．

5.1　算出手法

まず，それぞれの口コミのポジティブな特徴語の出現数をポジティブスコア，ネガティブな特徴語の出現数をネガティブスコアと定義する．ここでの特徴語は表10，表 11 に示したそれぞれ 17 語を用いている．

このポジティブスコアからネガティブスコアを減じたものを各口コミの感情スコアと定義し，この値が正の値であればポジティブな感情の口コミと判定し，逆に負の値であればネガティブな感情の口コミとする．ゼロの場合はいずれにも属さないと判定する．このとき，評価 4 以上の口コミに対して，ポジティブな感情の口コミであると判定できれば正解とする．同様に，評価 2 以下の口コミに対し，ネガティブな感情の口コミと判定すれば，こちらも正解となる．このような形で正解率を求めた．

5.2　検証結果

評価 2 以下の口コミ 27 件に対する感情スコアに基づく判定の正解率を図 1 に，評価 4 以上の口コミ 17 件に対する正解率を図 2 に示した．図 1 から，評価 2 以下の口コミで感情スコアが負の値となったものはおよそ 80%となり，図 2 においても正解率は 6 割を超えた．

この結果から，ネガティブな特徴語として抽出した単語は，感情辞書に登録する単語として適切であるといえる．一方で，ポジティブな特徴語として抽出した単語も，感情辞書に登録する単語としてある程度機能するといえるが，不正解率が約 4 割あるため改善の余地がある．

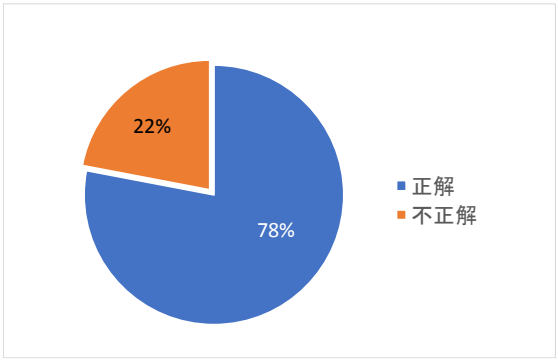

図1:評価2以下の正解率

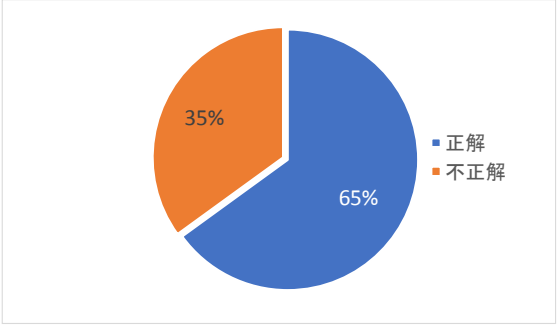

図 2:評価 4 以上の正解率

6.　終わりに

本研究ではグルメサイトの口コミに出現する単語について形態素解析し，ポジティブ・ネガティブそれぞれの出現頻度を単語の文章集合で除し，品詞ごとに出現率を示した．この出現率の高い単語を特徴語として，感情辞書に登録するポジティブ単語，ネガティブ単語として選び出すことができた．さらに，抽出したこれらの単語の適切さについて評価検証し，ネガティブな口コミをこれらの特徴語に基づく感情スコアによって，ネガティブであると正しく判定できる正解率が高いことも示し，感情辞書へ追加する単語として適切であることを明らかにした．

しかし，ポジティブな口コミに対する感情分析の正解率は7割を切り，特徴語として適切であるかどうかは判断しづらい結果となった．そのため，今後口コミを増やすことで特徴語を見直し，正解率の精度向上を図ることが必要である．

参考文献

[1]　気持ちが理解できる『感情分析』, http://www.textmining.jp/curation/emotion_lp.html

[2]　小林雄一郎，Rによるやさしいテキストマイニング，オーム社，2017

[3]　三和未佐希，立間淳司，　青野雅樹，“単語位置と強弱表現に着目したツイートの感情分析”，FIT 2014，pp. 227-228，2014

[4]　小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一，“テキストマイニングによる評価表現の収集”，情報処理学会研究報告，154(12)，pp.15-20，2003，

[5]　村上菜緒，尼岡利崇，牛田高信，木本勝敏，“twitter上で任意の検索語句に対するネガポジ度を判定し可視化するアプリケーションの開発と研究”，EC2014，pp.261-265，2014

[6]　石田基弘，R によるテキストマイニング入門，森北出版，2017

[7]　小町 守監修，奥野 陽，グラム・ニュービック，萩原正人，自然言語処理の基本と技術，翔泳社，2016